Policy Pilots and Evaluation

History and Policy Seminar April 2013

Centre for History in Public Health with the Policy Innovation Research Unit LSHTM

Summary report

KEY MESSAGES

- Key concepts such as *pilots* and *what works* need to be carefully defined and refined
- Pilots have different implications within different models of evaluation
- Interest in policy piloting has grown over time, especially under New Labour administrations, with increasing emphasis given to rigorous (outcome) evaluation, preferably involving RCTs
- The impact of pilots and evaluation is affected by context and the complex nature of the policy process
- Expectations that evidence from pilots might translate into policy in a straightforward manner have often been disappointed, partly due to the difficulty Governments have encountered in accepting findings that did not confirm a given policy direction
- There are also other reasons for this reluctance, such as where findings from piloting have been too narrowly focused on a particular case of policy implementation
- Over time, Governments seem to have become more risk averse in a context of close and extensive media scrutiny
- Paradoxically, small, low profile pilots may be more successful in generating learning 'under the radar' than larger, high profile priority programmes
- Another paradox is that Governments have often been enthusiastic in talking about and instigating piloting and evaluation but have been less keen to use the findings from such evaluations
- There are multiple audiences for evaluation results – including the general public, NGOs and service users: researchers should be mindful of these, not only governments and commissioners.

Policy Pilots and Evaluation

Health History and Policy Seminar
April 2013

Key Themes

- The role of evaluation and evidence in policy making
- Different conceptualizations of the policy process
- Continuities and changes over time
- Different understandings of what is a pilot
- Uses and misuses of pilots and/or evaluation findings
- Different cultures of government departments
- Influences and constraints on the usefulness of pilots
- Ministers, politics and evidence
- Interests of different stakeholders
- Methodological issues
- International comparisons

Stefanie Ettelt Policy Innovation Research Unit LSHTM

Health policy piloting in England in the 2000s: has the drive towards 'evidence based policy' resolved long-standing dilemmas in relation to the purpose of piloting? - Stefanie Ettelt and Nicholas Mays, LSHTM.

Stefanie Ettelt based her analysis on policy pilots initiated by the Department of Health (DH) in the wake of the publication of the 2006 White Paper *Our health, our care, our say*. The context for these was the commitment of the then Labour Government to evidence-based policy-making and 'what works' (HM Government, 1999). The talk revisited an argument made by Martin and Sanderson (1999) who had noted (reflecting on the introduction of the *Best Value* pilots in the late 1990s) that these pilots were introduced to meet several objectives simultaneously, such as to measure impact, identify good practice and facilitate learning for implementation. The authors wondered whether one study design can meet such a diverse set of expectations of evidence, given that each approach has different theoretical, practical and methodological implications.

Ms Ettelt noted that in her study – which examined the *Partnerships for Older People Projects (*POPP) pilots, the *Individual Budgets* pilots and the *Whole System Demonstrators* – a similarly diverse set of purposes of piloting could be identified:

- Piloting for experimentation, which is most in line with the aspirations of evidence-based policy
- Piloting for early implementation, i.e. as an opportunity to facilitate local change (as the term 'pioneer' which is sometimes used, seems to imply)
- Piloting for demonstration, to show others (not involved in the pilots) how best to implement a policy
- Piloting for learning how to operationalize policy (as the terms 'trailblazer' or 'pathfinder' suggest).

It was notable that there was an increased interest in policy piloting under the New Labour Governments, specifically between 1999 and 2008 (i.e. between the publication of the 1999 White Paper *Modernising Government* and the 2007/8 fiscal crisis). A key reference for piloting is the 2003 Cabinet Office report *Trying it Out* (Cabinet Office, 2003). This report had proposed that an ideal form of piloting would involve the 'rigorous early evaluation of a policy (or some of its elements) before that policy has been rolled out nationally'. Adjustments to the proposed programme would then be made in the light of evidence from the pilot. Using this definition, the report found that many pilots were not strictly 'pilots', as many were not evaluated or findings from evaluation did not have an influence on policy decisions, as decisions were made before evaluators reported their findings.

So there is a paradox: on the one hand, there was substantial enthusiasm by Government officials for piloting and for using pilots as policy experiments; on the other hand, there was a lack of engagement with the findings of pilot evaluation, especially of those findings that were not as supportive of policy as perhaps expected by those initiating the pilots.

The inability of governments to tolerate an open outcome and accept genuine uncertainty as stipulated in the idea of experimentation resonates with earlier observations by Donald Campbell (1969) who had noted that governments tend to commit to policy politically and thus find it difficult to be seen at fault. In his view, this undermines their ability to learn from experience, especially to learn from failure.

A similar dynamic was observed in two of the three pilot programmes analysed here, which involved randomised controlled trials (RCTs). The Treasury, particularly in this period, had a keen interest in RCTs, seen as at the top of a 'hierarchy of evidence', to provide solid evidence of outcomes. These, combined with measures of cost-effectiveness, would then provide insights into whether the policies 'worked' and how

much they would cost. Ms Ettelt pointed out that, in the context of the realities of the policy-making process, this is hugely ambitious and problematic, as an assumption was made at the outset that the policies worked and that the DH could therefore support its bid for resources to the Treasury with such evidence. It was also notable that the Treasury had a keen interest in the pilots at the onset, but there was no evidence of any sustained engagement in the research or in responding to the findings.

The pilots also involved a number of different stakeholders, with different interests and agendas, all of whom hold different (and competing) expectations of the pilots. Tensions between piloting for experimentation and piloting for implementation could be seen in each of these three cases.

Another contextual feature observed by Ms Ettelt in her research was that interest seemed to change at the DH over time, thus shifting the purposes of the pilots. In one case, this involved a shift from piloting for early implementation and innovation to measuring outcomes, with a strong (and sudden) emphasis on methodological rigour which had not been anticipated at the planning stage of the pilots. However, these shifts in purposes played out differently in each case, suggesting that there was not one direction of development (e.g. towards more rigorous evaluation), but an inherent ambiguity of purposes that characterised these policy pilots. These ambiguities made it difficult for researchers to produce evaluations that were relevant to policy-makers.

Philip Davies [Deputy Director and Head of the London Office of 3ie]

What we learned from the 2003 Review of Policy Pilots in the UK Government
Philip Davies, PhD

Phil Davies had played an active part in the report *Trying it Out* referred to by Stefanie Ettelt. He observed that pressures for change emerged not only from the New Labour Ministers, especially the Prime Minister, but also from the Civil Service.

A key feature of the background here was the *Modernising Government* White Paper (HM Government, 1999), the aim of which was to use evidence to make more informed decisions and get better policy making for the 21st Century. Dr Davies also referred to the two types of pilots: for experimentation and for implementation. In the first type, the aim would be to collect evidence on the effects of policy change which could be tested against a genuine counterfactual. This view thus privileged methodologies based on

the randomised controlled trial. This research designs favoured in medical research, which were thought to give the strongest possible evidence of what works. Proponents of this view felt that if social policy experiments could be constructed on these principles better outcomes would ensue. Dr Davies noted, however, that there are many examples of bad randomised controlled trials.

Supporters of the other approach to piloting – focusing on processes and concerned with issues of implementation – observed realistically that the conditions for experimentation often did not exist: especially this was because Ministers would often announce a pilot and at the same time announce its national implementation and roll out. Some evaluations of a pilot might show that the Government's policy did not work, but they had proceeded with it anyway. Dr Davies also mentioned 'phased implementation pilots', whereby policy would be implemented in stages, and to the extent that the policy could be supported by existing evidence. Where there was uncertainty about the effectiveness of a policy, or about the best way to implement it, further piloting would be undertaken.

The question 'what works' also needs to be refined: evaluation needs to ask for whom it works and at what costs does it work, and under what conditions? There are a number of important methodological issues that need to be thought about carefully in this field, including the question of when to take action based on reported outcomes? Early impacts may not be sustained over time, and lack of outcomes in the short term might be followed by positive outcomes later.

When one large-scale RCT was undertaken, the contract went to an American contractor, because UK civil servants were not satisfied that sufficient capacity existed in the United Kingdom.

Dr Davies also mentioned the 'long grass theory of pilots': i.e. if you want to get rid of a problem politically, or from a policy point of view, 'do a pilot' - because it kicks the issue into the 'long grass' for a period of time.

Appreciation of the value of pilots might however come from realising the advantages of saving the Government or a Minister from embarrassment by spotting faults early on, before programmes had been extensively rolled out, or had attracted high costs. A problem, however, is that research timetables and policy timetables often do not coincide.

Martin Bulmer

The relationship between research and policy: Some particular considerations

Martin Bulmer, University of Surrey

Professor Bulmer spoke on some general issues about the relationship between research and policy. Lying in the background is the question how far one can learn from the processes of evidence-based medicine and extrapolate from these to support the argument that social policy should be based on evidence.

He reminded the audience of the two models *engineering* and *enlightenment* (Weiss, 1979) and indicated his preference for the enlightenment model as a more accurate portrayal of the way in which knowledge can influence policy. The engineering model has a rather too simplistic idea of how far it is possible for policy to bring about behavioural change. Central to the discussion is the question of how policy-makers use the knowledge that is produced. It is also necessary to retain a sceptical view on the soundness of the scientific analysis underlying any piece of social research. It is essential also to understand the policy context in which a pilot is being carried out.

A fundamental question for policy roll-out relates to the extent to which one can generalise from the results of any one pilot case study to a larger class of cases.

Professor Bulmer concluded that in his view it is very difficult to achieve social interventions on the model of the systematic review of evidence that has been so influential in the medical field.

DISCUSSION

Past and recent history

In the discussion that followed a number of interesting observations were made, drawing on the expertise of (sometimes former) civil servants and researchers present. The periods discussed were mainly within UK: the 1960s; 1970s; the early, mid and late Blair administrations; the Brown years; and the current Coalition Government. While the emphasis was on government-funded pilots, an important observation was the need to extend the discussion beyond Whitehall to include local government, devolved administrations and other 'outside' groups and organisations: these may engage with the findings from pilot evaluations even if the Government chooses to ignore them.

Illustrative examples were discussed from a number of pilots, including pilots of the Educational Maintenance Allowance, Welfare to Work, Sure Start, Drug Treatment and Testing Orders, Community Pharmacies, Screening programmes, Patient Reported Outcome Measures (PROMs), and Best Buy and Best Value projects. Most of the pilots cited during the seminar had taken place in the recent past and in particular under the

Labour government after 1997. Other pilots referred to were in the 1960s, 70s and 80s, such as Educational Priority Areas, Community Care and the Home Office Urban Programme. Whether these were significantly different from the later pilots was debated.

Contextual influences

It was agreed that the context for policy pilots was often highly complex and highly politicised.

Many of the observations came from policy pilots and their evaluations conducted during the New Labour administrations. An important historical question raised was whether there was something particular about the New Labour years, which were characterised by substantial growth in public spending (between 1999 and 2007), but also great uncertainty about the direction of travel for policy. It was argued that the emphasis on evidence based policy attempted to fill an ideological gap within New Labour.

It was generally agreed that the time frames for piloting have never matched up with the political time frame, with one participant stating that 'by the time you've actually reported the findings from your pilot, you're almost bound to be working on the next policy'.

Flaws in the process of pilot evaluation may have come from over optimism at the outset of piloting, weak methodologies and a thirst for policy to be seen as success. In the 2000s, in some government departments, there seemed to be a growing intolerance of bad news. This could put pressure on evaluators to produce welcome findings.

Politicians and evidence

Several participants recalled incidences when politicians seemed to completely ignore all evidence.

Partly this reflects the different time horizons of politicians and researchers: 'however good your pilots, however good your evaluations, if you have politicians with short time horizons, they are only likely to accept your results if they give them good news within those time horizons'.

Even where results from a pilot seemed to have impact on policy, findings from research often appeared to have played little part in the final decision. As one researcher involved in evaluation recalled:

'One version of history would be that this pilot shaped the policy because this is what the sequence of events suggests. But before we'd even finished our report and presented the findings, the policy had been made at the highest level in the Department without any apparent reference to the little pilot researcher'.

Evidence from an ongoing pilot may percolate through a government department. One former civil servant suggested that 'if a pilot was going to work you knew it was working before the research report was delivered'. This links to another observation, that the evaluation and the evaluator are themselves actors in the policy process, influencing patterns of change: this influence may be direct or indirect.

It was felt however that it is not surprising that Governments do things because they *believe* in them, rather than on the basis of evidence, given that they are typically elected for the ideas and values they stand for.

A question was raised as to whether politicians were actually the main problem: in earlier years, opposition to evaluation had been observed to have come mainly from civil servants, even though they might have pretended there was opposition from Ministers.

This view was supported by another participant who observed that there is a predilection among some civil servants to use personal networks rather than documents as a source of knowledge, suggesting that using evidence might still conflict with civil servants' ways of working. However one participant observed that some trends are encouraging. Government officials now use terms such as *theory of change* or a *logic model*, indicating that there are attempts to explicitly think about the mechanisms, models and activities that are expected to produce policy outcomes. Terms like *benchmarking* and *best practice* have also come into their vocabulary.

A civil servant felt there were some positive developments currently, for example, public officials are being encouraged to acknowledge failures and write up lessons learned. Some of the apparent unwillingness to countenance failure comes from the pressures of the media, with commentators often being intolerant of politicians reversing previous decisions. In this country, U-turns are often seen as a weakness of leadership rather than recognition of learning.

A related view was that many government departments now have strategy units and in this context, it is sometimes possible to conduct pilots and research that take a more long-term perspective, i.e. beyond the next spending cycle.

Another more optimistic comment was that 'smaller', less high profile pilots are often more likely to influence policy decisions and effect changes, as they may develop 'under the radar'. The hope is that over time a series of incremental changes might add up to a larger change without attracting the same amount of political or media attention. This way it may be possible to learn from

failure as well as success. Other participants agreed that evidence is more likely to be used for decisions that are 'mundane' and not highly charged ideologically.

There were diverging views about whether evidence use in health policy is different from in other areas of social policy in England: one view was that differences between policy areas may in practice be quite small.

The purposes of pilots

Adding to the discussion of types of pilots, it was observed that sometimes pilots are introduced because resources are tight and only a pilot can be afforded at that time. Pilots might also be seen as a tactic for enrolment, for getting the support of professionals or particular stakeholder groups, who may be less likely to object to a policy proposal if it is presented as a pilot.

Yet experience also suggests that pilots can be used by policy-makers to legitimise a course of action and to add to their authority. This may be especially the case in relation to policy decisions where politicians are dependent on the support of others to facilitate change or where their ability to influence change directly is limited.

More in line with the idea of 'evidence based policy' was the observation that some pilots were indeed used to establish the evidence base for policy, providing a valuable 'challenge function'.

Methodological issues

There were some concerns that British social research was not of a high enough quality to evaluate large scale interventions: America had for some time been using RCTs in social policy, which are seen by some as the highest standard of research evidence.

It was questioned, however, whether social policy pilots could ever be set up as experiments. Even if the evaluation uses an experimental research design, the premise of experimentation – that genuine uncertainty exists about the outcome of an intervention – usually does not hold in social policy. One participant noted that even if experimentation was the intended purpose at the beginning of a pilot, policy-makers were unlikely to tolerate uncertainty during the course of a pilot, thus putting pressure on evaluators to produce findings in support of the policy.

The lack of clarity among practitioners and civil servants as to the aims of a project can present problems for researchers. Participants reported that it often is not entirely clear what is being piloted. This lack of clarity may be revealed in the process of evaluation. This may indicate differences among participants about the goals of a pilot. From an evaluation perspective, such ambiguity of purpose is not helpful, even if

one concedes that different audiences may have different interests in and aspirations for pilots and their evaluations.

It was also noted that, from a policy perspective, the fact that evaluation of pilots generates findings that are specific to these pilots can be a disadvantage. Researchers should therefore seek to ensure that findings can be useful more broadly. However, it was also noted that there are often limits to generalisability, although researchers may tend to define these limits more narrowly than policy-makers. Contributing to an evolving evidence base may be a way forward to ensure that evaluation is useful for policy beyond decisions immediately related to the pilot and the policy it refers to.

There seem to be moves towards accumulating such evidence at European level, linked to the use of the open method of coordination. Such evidence bases require systematic and continuous institutionalising of monitoring practices. A plea was made to use evidence from research in Europe rather than always be looking to the USA for evidence and ideas.

Related to this was the observation that while the spotlight is often on RCTs, qualitative research was having an impact on policy and bringing about innovations. Qualitative evaluation methods are now much more recognised as having a legitimate and valuable contribution to make to answering more complex evaluation questions about 'what works, where, when and for whom'.

Participants also endorsed the importance of data on cost-effectiveness. The growth of influence of health economists was pronounced in the DH and in government more generally, with one participant stating that health economics in the past 'came in like an express train'. At the DH, for example, the number of economists went from six in 1974 to over 50 by the late 1990s.

As new disciplines came into government, different methodologies came to prominence: modelling and the use of large data sets, such as the Family Expenditure Survey or the British Household Panel Survey had a substantial influence on social security policy. Using these methods allowed policy-makers to quantify the impact of policy in a way which had previously been impossible.

Audiences for results and findings

An important comment was that it is a mistake to assume that policy 'customers' who have commissioned a piece of research are the only, or perhaps even the most important, users of pilot evaluations. Experience suggests that there are multiple audiences, including those that use evidence

in the longer term such as interest groups advocating for or against policy. These alternative channels of evidence-influence may be more influential in the long run than the more immediate 'users' of policy i.e. 'customers'. Pilots initiated by the Young Foundation or Joseph Rowntree Foundation were mentioned as examples of initiatives that take place outside Government. These might be small scale pilots involving citizens or public participation and different kinds of research methods. It was also observed that significant cultural differences exist between different government departments with some more open to the use of evidence than others.

CONCLUSIONS

Overall there seemed to be a sense of pessimism and disappointment with the way policy pilots and evaluations are currently used and were used in the past, and with the limitations of the policy process to engage with findings. Influences identified included: poorly designed studies; weak methodologies; impatient political masters, time pressures and unrealistic deadlines; persistent, some say even growing, intolerance of bad news and pressure on evaluators to 'say the right thing', leading to an erosion of trust between evaluators and officials; the media contributing to the risk aversion of politicians with their tendency to value 'decisive' Ministers, while a policy U-turn (in response to negative findings) is presented as a failure; and a lack of incentives for civil servants to become involved in 'riskier', experimental initiatives, paired with an over-reliance on being seen to be in charge of a policy 'success'. However, some more positive developments were noted as well: examples of beneficial institutional change, supportive of the use of evidence, included the National Institute for Health and Care Excellence (NICE) or the National Institute of Health Research (NIHR), which seem to pursue a longer-term agenda of building evidence bases in some areas of health policy. Reference was also made to lessons that might be learnt from governments in other countries such as in Europe and some saw promise in the initiation of the new *What Works Centres*.

The discussion revealed a central paradox: findings from evaluations of pilots should be specific to allow learning to improve the policy, but they should also be generic enough to be applicable to other situations. However, there are typically severe limitations to the generalisability of findings from evaluation of policy pilots. These limits relate to the nature of pilots and the fact that effectiveness of policy is often highly context dependent. Pilots by definition only cover part of this context, as they are limited geographically (e.g. involving only a selection of local areas) and in terms of time. There are therefore questions to be answered if findings from such evaluations are to be interpreted to make recommendations for 'scaling up', policy roll-out or policy transfer.

LIST OF PARTICIPANTS

Virginia	Berridge
Nick	Black
Annette	Boaz
Martin	Bulmer
Ross	Coomber
Tony	Cutler
Philip	Davies
Lucy	Delap
Bob	Erens
Stefanie	Ettelt
David	Gough
Linda	Hantrais
Ben	Hawkins
Paul	Hayes
Walter	Holland
Mike	Hough
Peter	Howitt
David	Lawrence
Janet	Lewis
Susanne	MacGregor
Frederick	Martineau
Nicholas	Mays
Gareth	Millward
Alex	Mold
Glen	O'Hara
Jane	Seymour
Peter	Shapely
Nicola	Singleton
Chris	Sirrs
Clive	Smee
Alison	Tingle
Liz	Toon
Sierra	Williams

THE SEMINAR SERIES

The series of seminars 2013-14 are organized jointly by the Centre for History in Public Health at LSHTM, University of London, and the Wellcome Centre for the History of Medicine at the University of Manchester. The seminar series arose from the interest of historians in having an impact on health policy and policy more generally. The idea behind the seminar series is to bring together historians, social scientists and policy people to discuss some of the key issues of the day. Other topics in the series include Cancer Screening and Alcohol Policy.

This seminar on Policy Pilots and Evaluation was arranged by the Centre for History in Public Health along with colleagues from PIRU (Policy Innovation Research Unit, led by LSHTM and funded by the Department of Health). The first session was chaired by Professor Virginia Berridge, Director of the Centre for History in Public Health. The second session was chaired by Professor Susanne MacGregor, Professor of Social Policy at LSHTM. Organisational support was provided by Alex Mold and Ingrid James. The seminar was funded by the Wellcome Trust.

References

Cabinet Office (2003). *Trying it out: The role of 'pilots' in policy-making: Report of a review of government pilots*. London, Cabinet Office, Strategy Unit.

Campbell, D.T. (1969) Reforms as experiments. *American Psychologist*, 24, 409-429.

H.M. Government (1999): *Modernising Government White Paper*. London, The Stationary Office.

Martin, S. and Sanderson, I., (1999), Evaluating public policy experiments. Measuring outcomes, monitoring processes or managing pilots? *Evaluation* 5 (3): 245-258.

Weiss, C.H. (1979): The many meanings of research utilization. *Public Administration Review* 39: 426-431.